Точные оценки вероятности переобучения для симметричных семейств алгоритмов

Фрей Александр Ильич

Московский физико-технический институт
(Государственный университет)
Факультет Управления и Прикладной Математики
Кафедра «Интеллектуальные Системы» (ВЦ РАН)

Научный руководитель: к.ф.-м.н. Воронцов Константин Вячеславович

22 сентября 2009

Проблема переобучения

- Строки таблицы $\{x_1 \ldots x_6\}$ — объекты полной выборки
- Столбцы $\{a_1 \ldots a_D\}$ — векторы ошибок алгоритмов

	a_1	a_2	$\ldots$	a_d	$\ldots$	a_D
x_1	0	1	$\ldots$	**0**	$\ldots$	1
x_2	1	1	$\ldots$	**0**	$\ldots$	0
x_3	0	0	$\ldots$	**0**	$\ldots$	0
x_4	1	1	$\ldots$	**1**	$\ldots$	1
x_5	1	0	$\ldots$	**1**	$\ldots$	0
x_6	0	0	$\ldots$	**1**	$\ldots$	0

- Метод обучения — минимизация эмпирического риска
- Цель: получить точные, вычислительно-эффективные оценки вероятности переобучения.

Содержание

1. Наделить структурой произвольное множество алгоритмов
   - Группа симметрий множества алгоритмов
   - Классы идентичных алгоритмов
2. Вывести общую оценку вероятности переобучения с учетом симметрии семейства
   - Рандомизированный метод обучения
   - Теорема о равном вкладе идентичных алгоритмов в вероятность переобучения
3. Исследовать семейства алгоритмов простой структуры
   - Явная формула для связки монотонных цепочек
   - Графики вероятности переобучения для трех случаев

Группа симметрий множества алгоритмов

- Генеральная выборка $\mathbb{X} = \left(x_i\right)_{i=1}^{L}$
- Алгоритм — бинарный вектор $a \equiv \left(a(x_i)\right)_{i=1}^{L}$ длины L
- Множество $\mathbb{A} = \{0,1\}^L$ — все алгоритмы длины L
- Аналогия:

Точка на плоскости	Алгоритм
Плоскость $\mathbb{R}^2$	Множество всех алгоритмов $\mathbb{A}$
Плоская фигура $F \subset \mathbb{R}^2$	Множество алгоритмов $A \subset \mathbb{A}$
Группа движений плоскости	Группа перестановок S_L

Определение (Группа симметрий)

Группой симметрий $S(A)$ множества алгоритмов $A \subset \mathbb{A}$ назовем его стационарную подгруппу:

$$S(A) = \{\pi \in S_L : \pi(A) = A\}.$$

Классы идентичных алгоритмов

- Орбитой элемента m множества M, на котором действует группа G, называется подмножество $Gm = \{gm\colon g \in G\}$.
- Две орбиты либо не пересекаются, либо совпадают.
- Разбиение на орбиты: $M = Gm_1 \sqcup Gm_2 \sqcup \cdots \sqcup Gm_k$.

- Группа $S(A)$ *действует* на множестве алгоритмов A.
- Алгоритмы внутри одной орбиты назовем *идентичными*.
- Обозначим $\Omega(A)$ — множества орбит, $\omega \in \Omega(A)$ — класс идентичных алгоритмов.

Теорема

Идентичные алгоритмы имеют равное число ошибок на полной выборке.

Рандомизированный метод обучения

- Метод минимизации эмпирического риска ставит в соответствие обучающей выборке алгоритм из заранее фиксированного множества: $\mu : 2^{\mathbb{A}} \times [\mathbb{X}]^{\ell} \to \mathbb{A}$

Определение

Рандомизированный метод обучения — это отображение вида

$$\mu : 2^{\mathbb{A}} \times [\mathbb{X}]^{\ell} \times \mathbb{A} \to [0, 1],$$

такое что для всех $A \subset \mathbb{A}$, $X \in [\mathbb{X}]^{\ell}$, $a, b \in A$ *и* $\pi \in S_L$ *имеем*

1. $\sum_{a \in A} \mu(A, X, a) = 1$;
2. $n(a, X) = n(b, X) \;\to\; \mu(A, X, a) = \mu(A, X, b)$;
3. $\mu(A, X, a) = \mu\big(\pi(A), \pi(X), \pi(a)\big)$.

Вероятность переобучения

- Вероятность получить алгоритм в результате обучения

$$P(a, A) = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \mu(A, X, a).$$

- *Вклад* алгоритма $a \in A$ в вероятность переобучения:

$$Q_\varepsilon(a, A) = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \mu(A, X, a) \big[\delta(a, X) \geq \varepsilon\big].$$

- *Вероятность переобучения*: $Q_\varepsilon(A) = \sum_{a \in A} Q_\varepsilon(a, A)$.

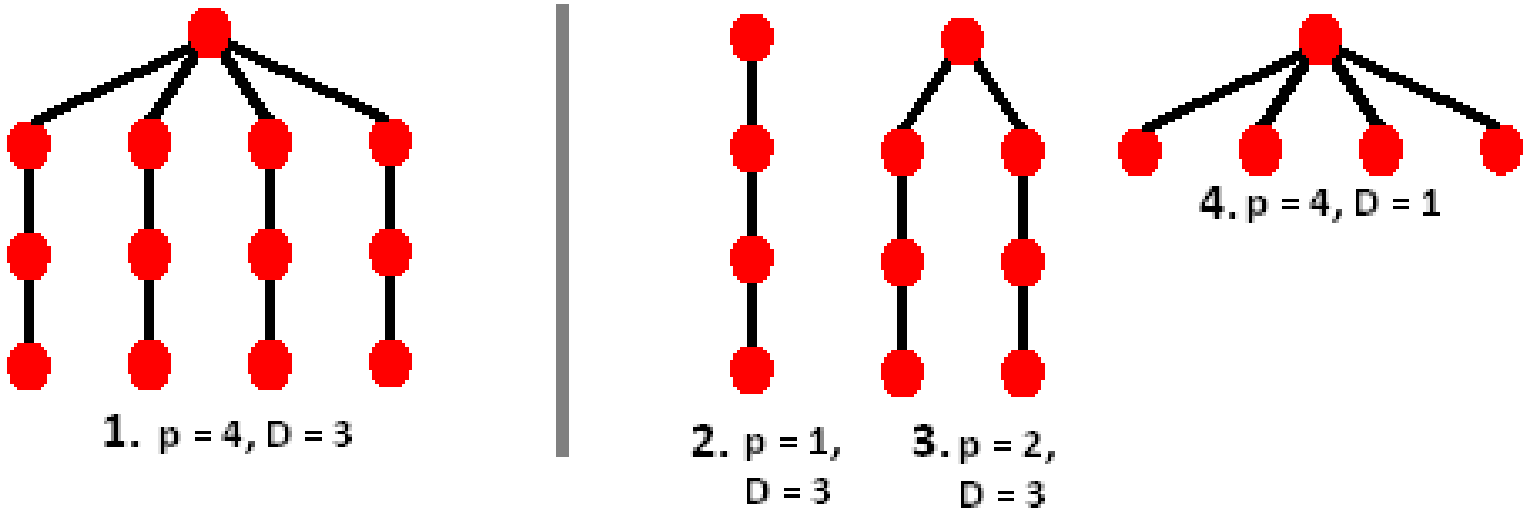
Теорема

Идентичные алгоритмы дают равный вклад в вероятность переобучения:

$$Q_\varepsilon(A) = \sum_{\omega \in \Omega(A)} |\omega|\, Q_\varepsilon(A, a_\omega).$$

Семейства простой структуры

- (1) Связка из монотонных цепочек
  - p — число ветвей
  - D — длина каждой ветви



- Частные случаи:
  - (2) Монотонная цепочка: $p = 1$.
  - (3) Унимодальная цепочка: $p = 2$.
  - (4) Единичная окрестность лучшего алгоритма: $D = 1$.

Явная формула вероятности переобучения

- Рандомизированный метод минимизации эмпирического риска.

Теорема

Для связки из p монотонных цепочек вероятность переобучения рандомизированного метода минимизации эмпирического риска может быть записана в явном виде:

$$Q_\varepsilon(A) = \sum_{h=0}^{D} \sum_{S=h}^{pD} \sum_{F=0}^{p} \frac{|\omega_h| R^h_{D,p}(S,F)}{1+S} \frac{C^{\ell'}_{L'}}{C^{\ell}_{L}} H^{\ell',m}_{L'}(s(\varepsilon)),$$

где $L' = L - S - F$, $\ell' = \ell - F$, $s(\varepsilon) = \left\lfloor \frac{\ell}{L}(m+h-\varepsilon k) \right\rfloor$; $|\omega_h| = 1$ *при* $h = 0$ *и* $|\omega_h| = p$ *при* $h \geq 1$, $H^{\ell',m}_{L'}(s)$ *— функция гипергеометрического распределения.*

Численные результаты - сравнение оценок

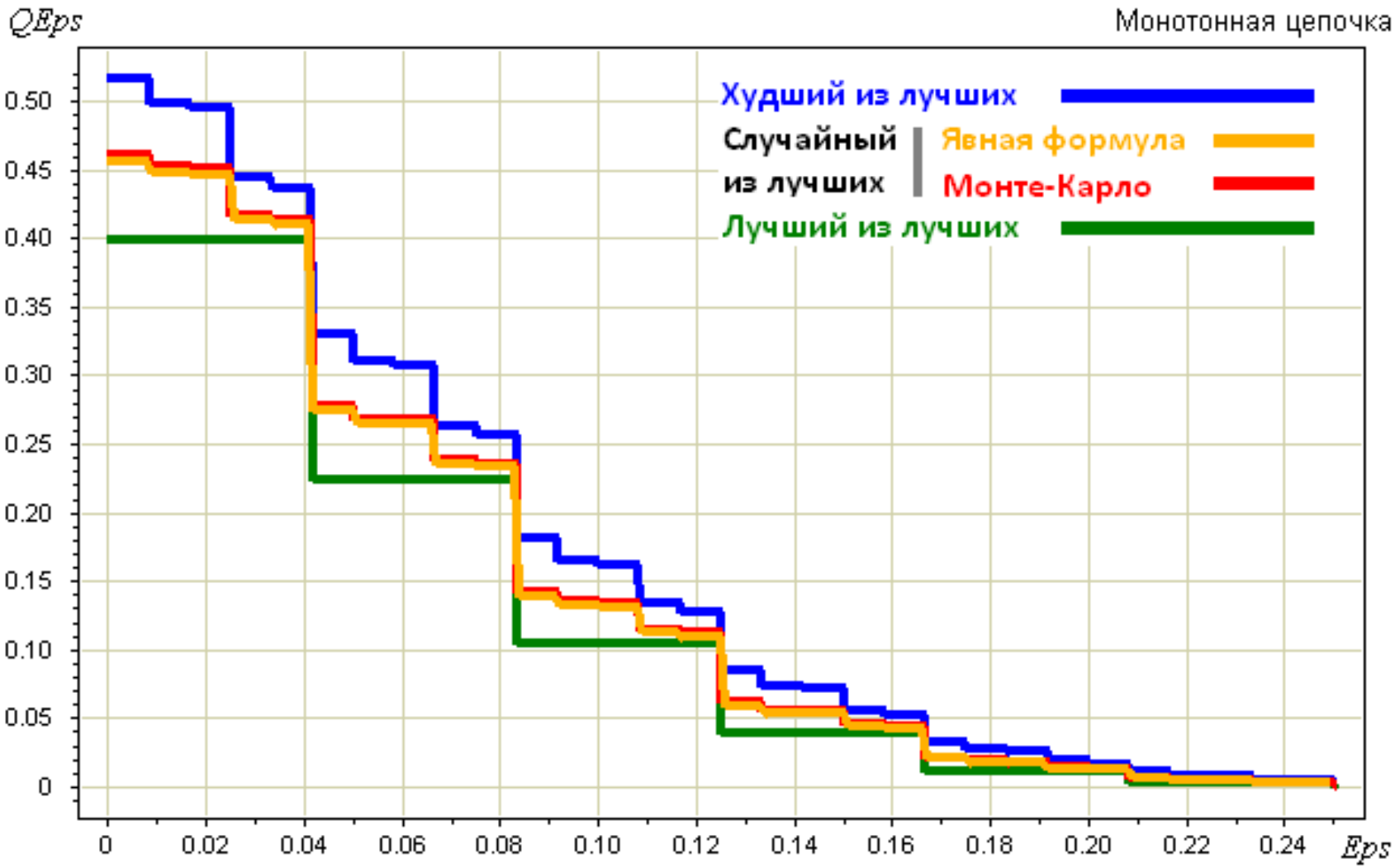


Рис.: $L = 100$, $\ell = 60$, $D = 40$, $m = 20$.

Численные результаты - сравнение оценок

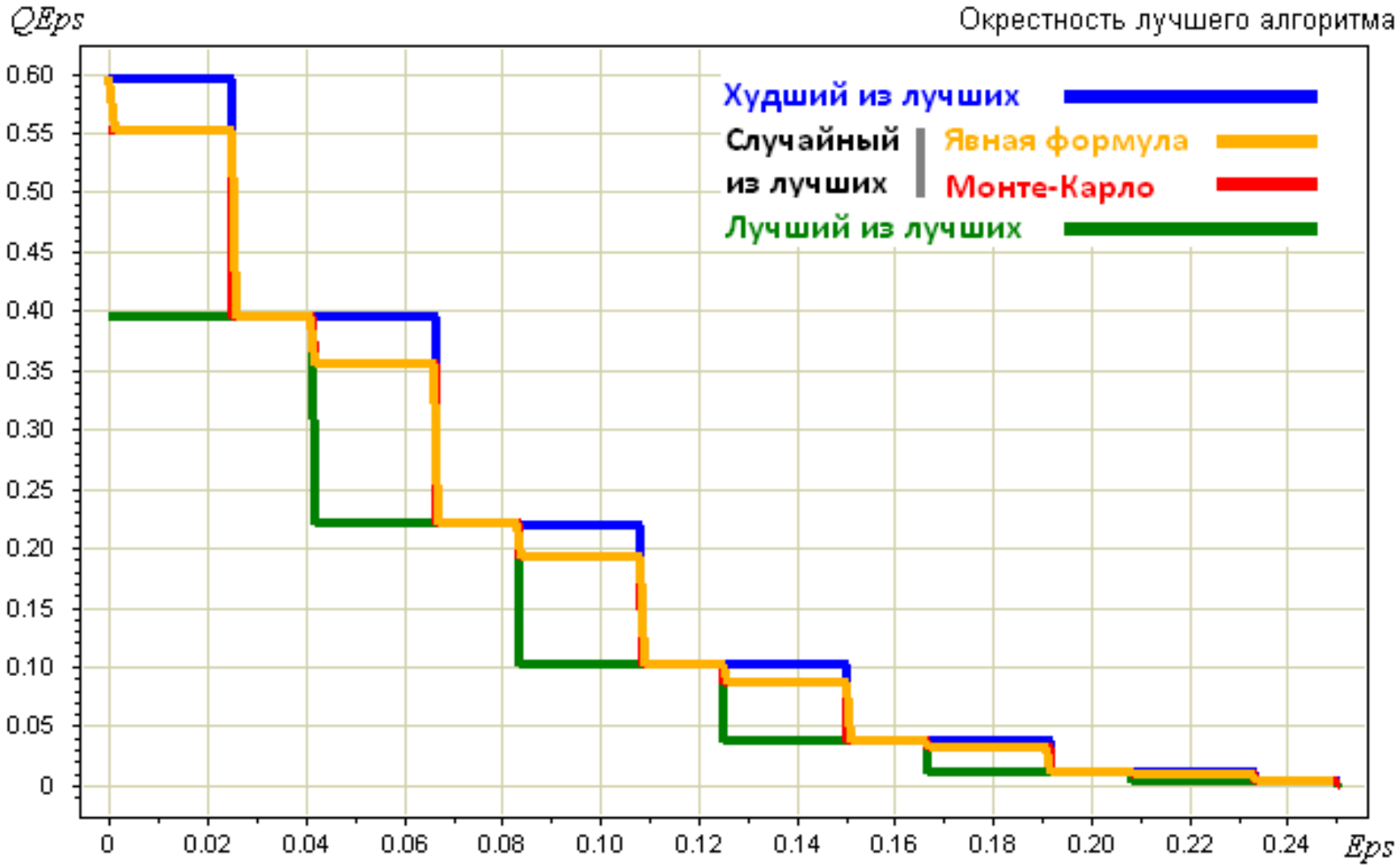


Рис.: $L = 100$, $\ell = 60$, $p = 10$, $m = 20$.

Основные результаты и направления исследований

1. Наделить структурой произвольное множество алгоритмов
   - Группа симметрий множества алгоритмов
   - Классы идентичных алгоритмов
2. Вывести общую оценку вероятности переобучения с учетом симметрии семейства
   - Рандомизированный метод обучения
   - Теорема о равном вкладе идентичных алгоритмов в вероятность переобучения
3. Исследовать семейства алгоритмов простой структуры
   - Явная формула для связки монотонных цепочек

Направления дальнейших исследований:

- Монотонные и унимодальные сетки большой размерности
- Исследование семейств алгоритмов, возникающих в реальных задачах